

FIG. 1

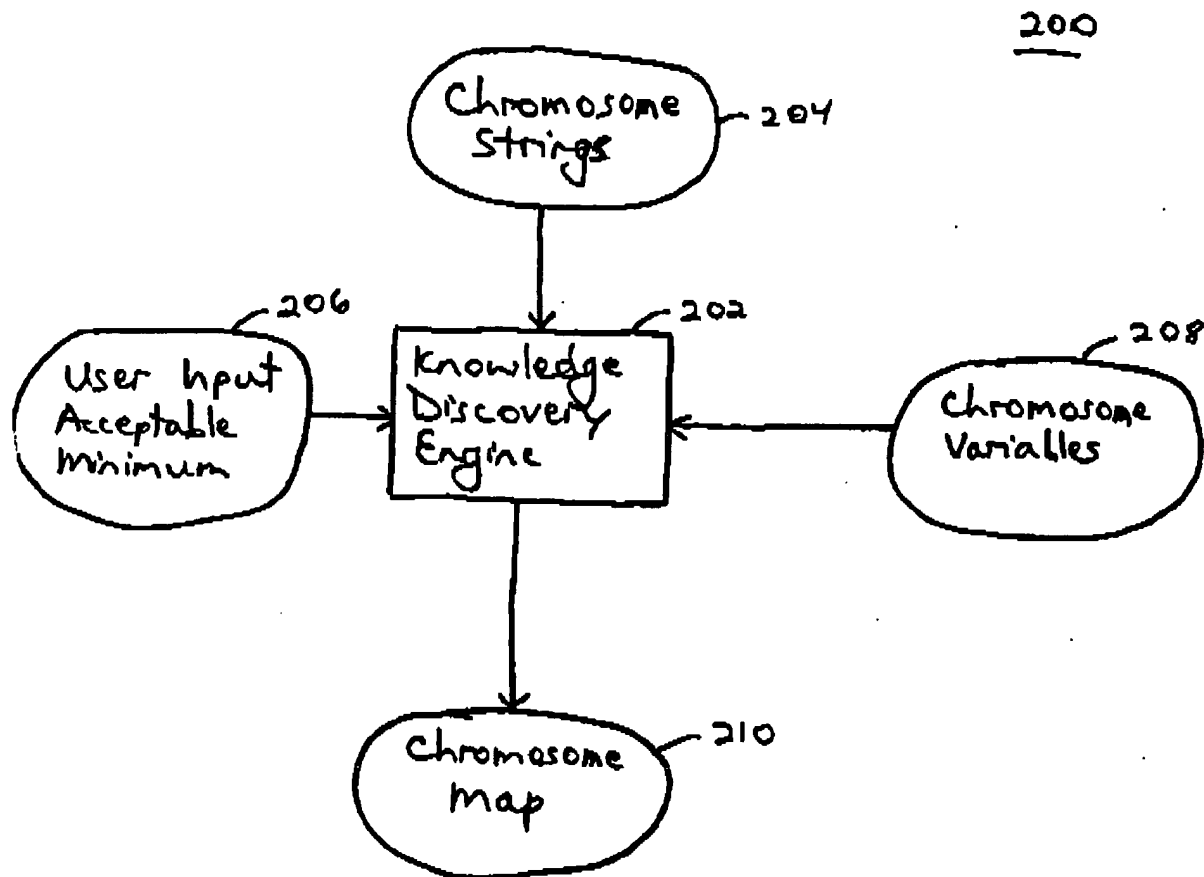


FIG. 2

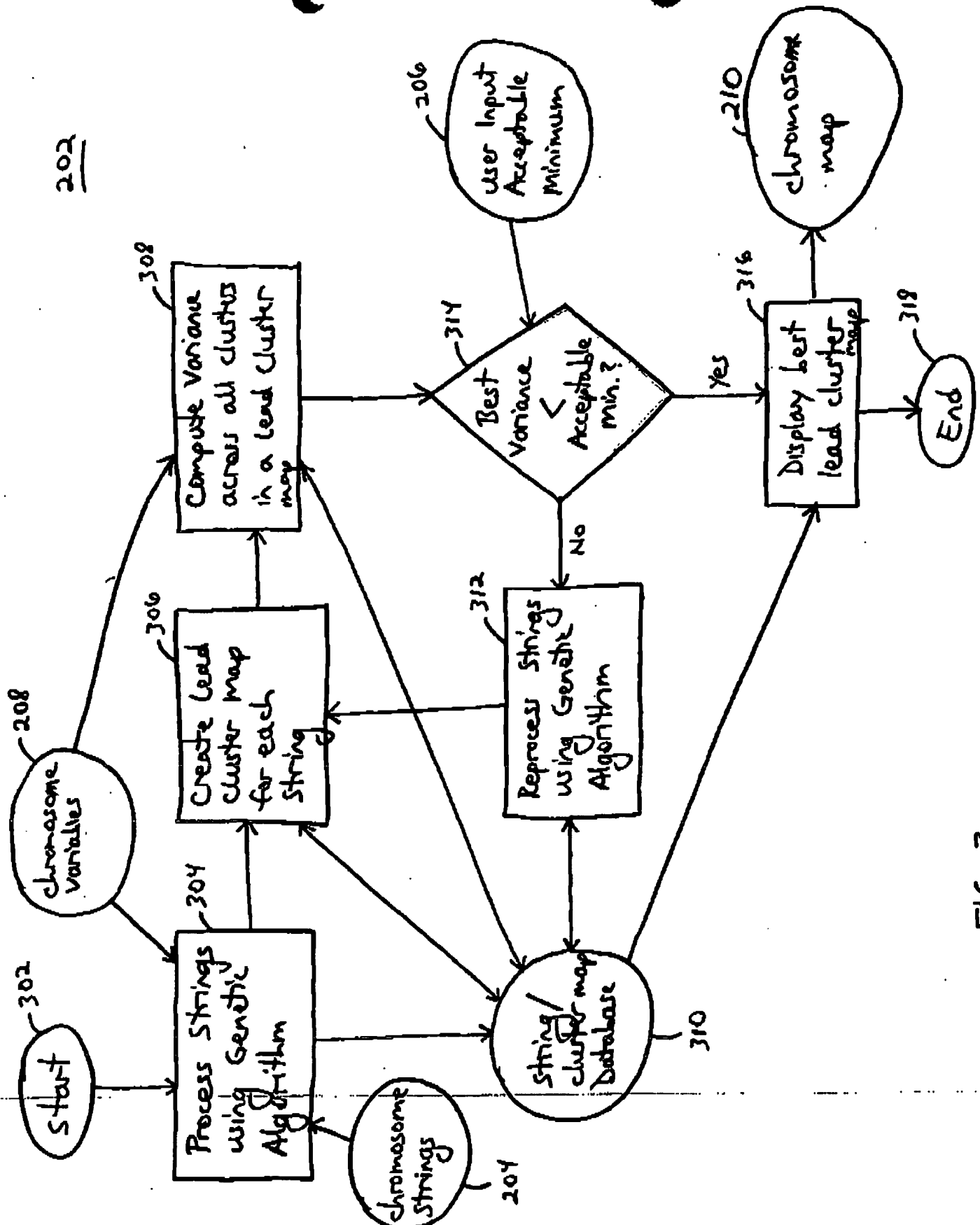


FIG. 3

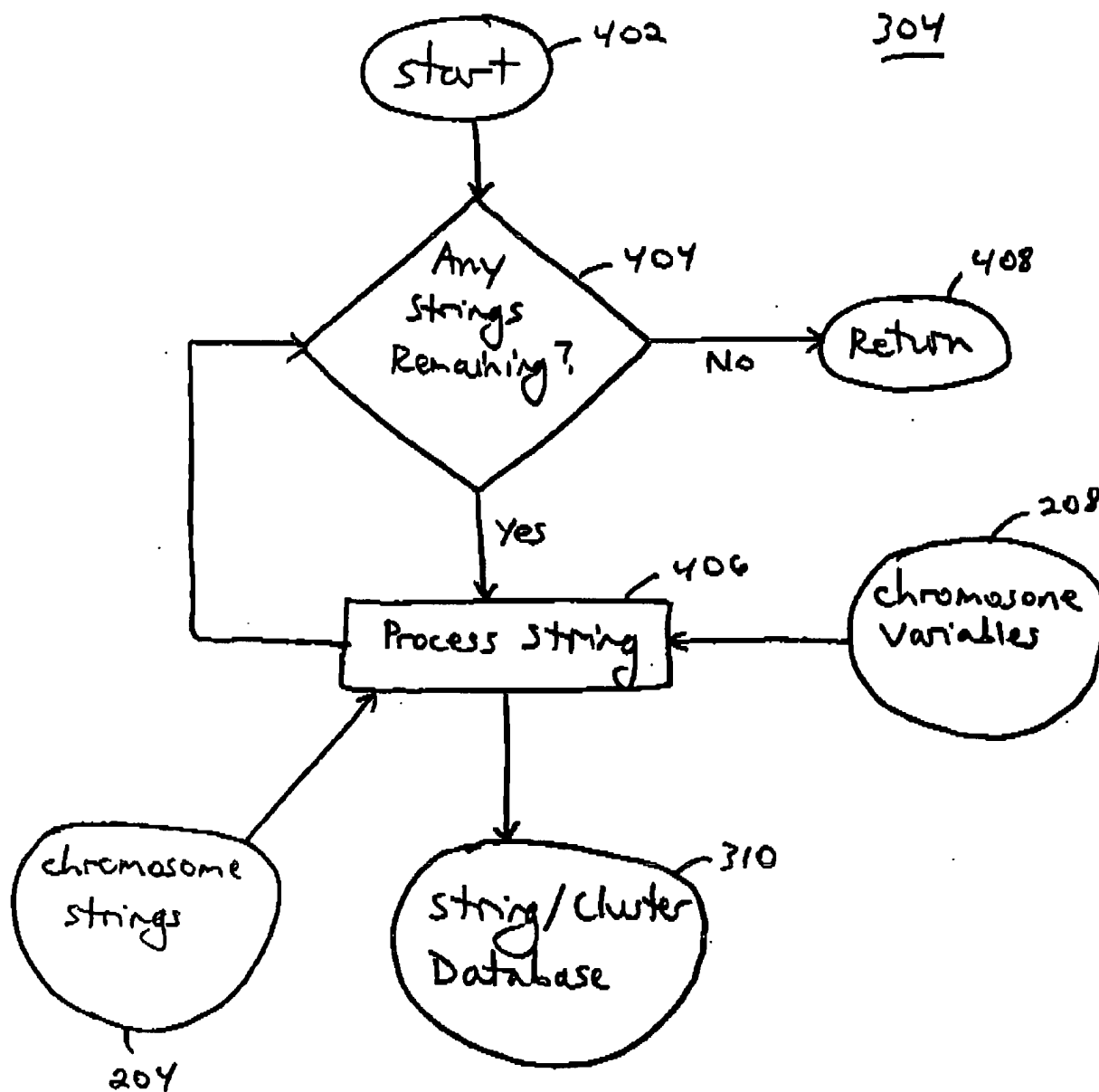


FIG. 4

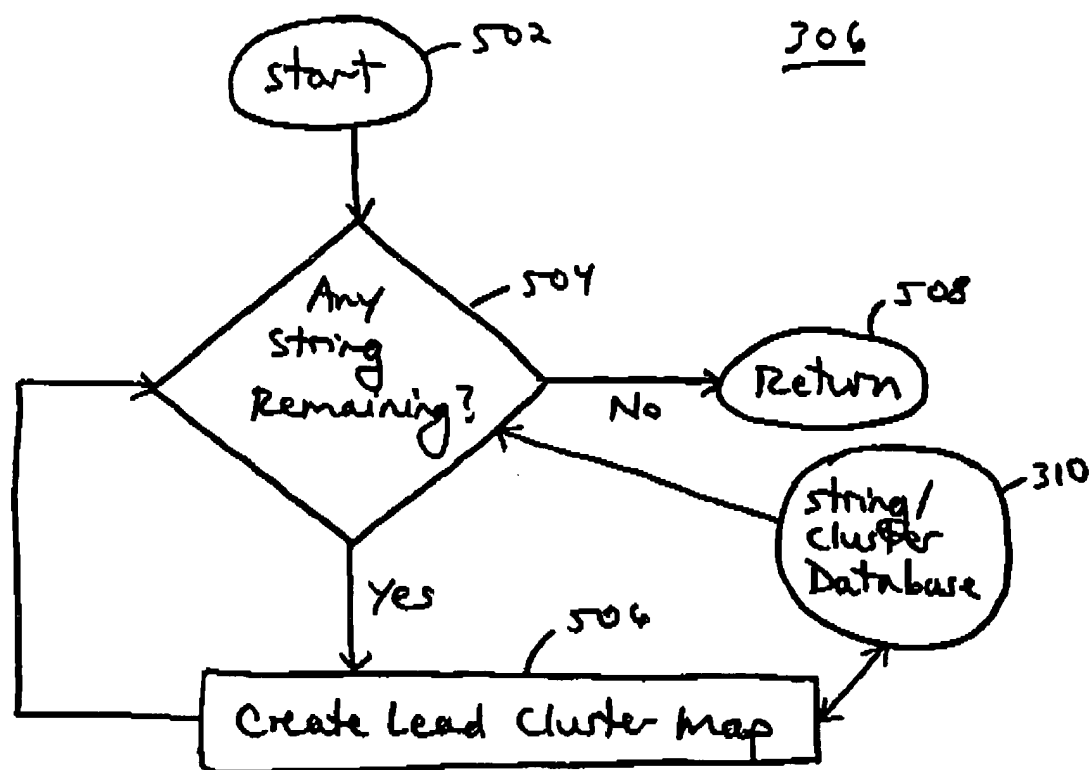


FIG. 5

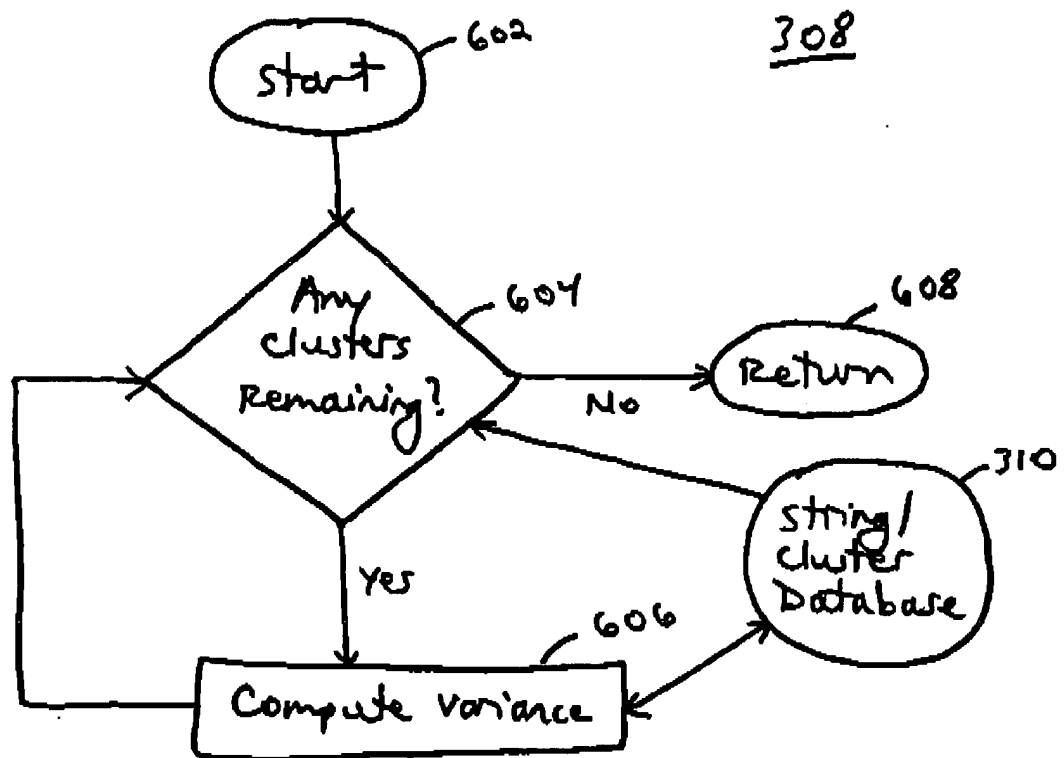


FIG. 6

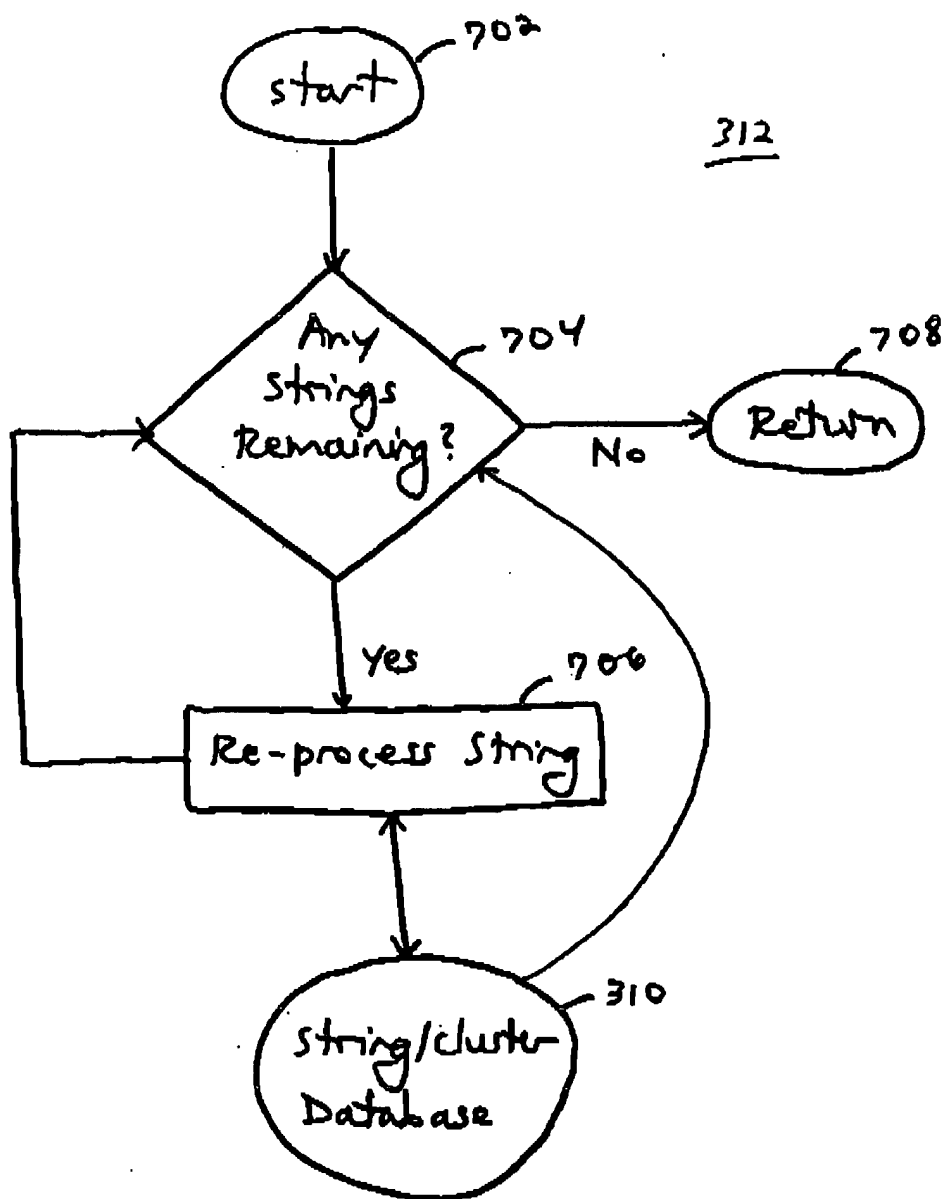


FIG. 7

## Knowledge Discovery Engine

- FIG. 3 is a control flow diagram showing the top level processing of the knowledge discovery engine. Processing begins at step 302 and immediately continues to step 304. In, step 304, the KDE 202 processes the chromosome strings 204 using a genetic algorithm. The chromosome strings 204 comprise data strings, e.g., bio-marker patterns, that are to be analyzed. The genetic algorithm inputs the chromosome strings 204 and for each data string, identifies the chromosome variables contained within the chromosome string 204. The chromosome variables 208 define the variables that the KDE 202 will look for in each chromosome string 204.

- The KDE 202 continues to step 306 and creates a lead cluster map, or grouping, for each processed chromosome string by using a pre-defined set of variables. The lead cluster map establishes clusters of data records around centroids in high order dimensional spaces. The membership of a record to a cluster is determined by Euclidean distance. If the Euclidean distance between a centroid and the record places the record inside a decision hyper-radius, the record belongs to the cluster surrounding the centroid. If the Euclidean distance between the record and any existing centroid is greater than the decision hyper-radius, the record establishes a new centroid and a new cluster. All data regarding the lead cluster mapping of the processed chromosome strings is recorded in the string/cluster database 310.

- The KDE 202 continues to step 308 wherein for each lead cluster map, it computes a variance across all of the clusters contained within that lead cluster map and records the variance in the string/cluster database 310. This step determines how homogeneous a given chromosome string 204 is to a predefined set of chromosome variable. The means for determining cluster homogeneity is a statistical measure of the variability of records belonging to a cluster with respect to specific behaviors, outcomes, attributes or the like.

- Upon completion of step 308, the KDE 202 determines a best lead cluster map; that is, it determines which lead cluster map is the "best fit" with the given set of chromosome variables.

- The KDE 202 continues to step 314 to determine whether the best lead cluster map is less than an acceptable minimum. The acceptable minimum may either be input by the user, or pre-defined within the KDE 202.



- If step 314 determines that the best lead cluster map is less than the acceptable minimum, then processing proceeds to step 316. In step 316, the KDE 202 records its final mapping in a chromosome map 210 and displays the best lead cluster map along with the matching variables.

- Returning to step 314, if the KDE 202 determines that the best lead cluster map is not less than the acceptable minimum, the KDE 202 proceeds to step 312.

- In step 312, the KDE 202 re-processes each processed chromosome string using the genetic algorithm. The genetic algorithm inputs the data for each processed chromosome strings from the string/cluster database 310 and reanalyzes them according to the last set of information. After completing the re-ranking of the processed chromosome strings, the KDE 202 returns to step 306 to create new lead cluster maps for each processed chromosome string. The processing continues as described above.

232127 v1/RE  
4z3z011.DOC

202

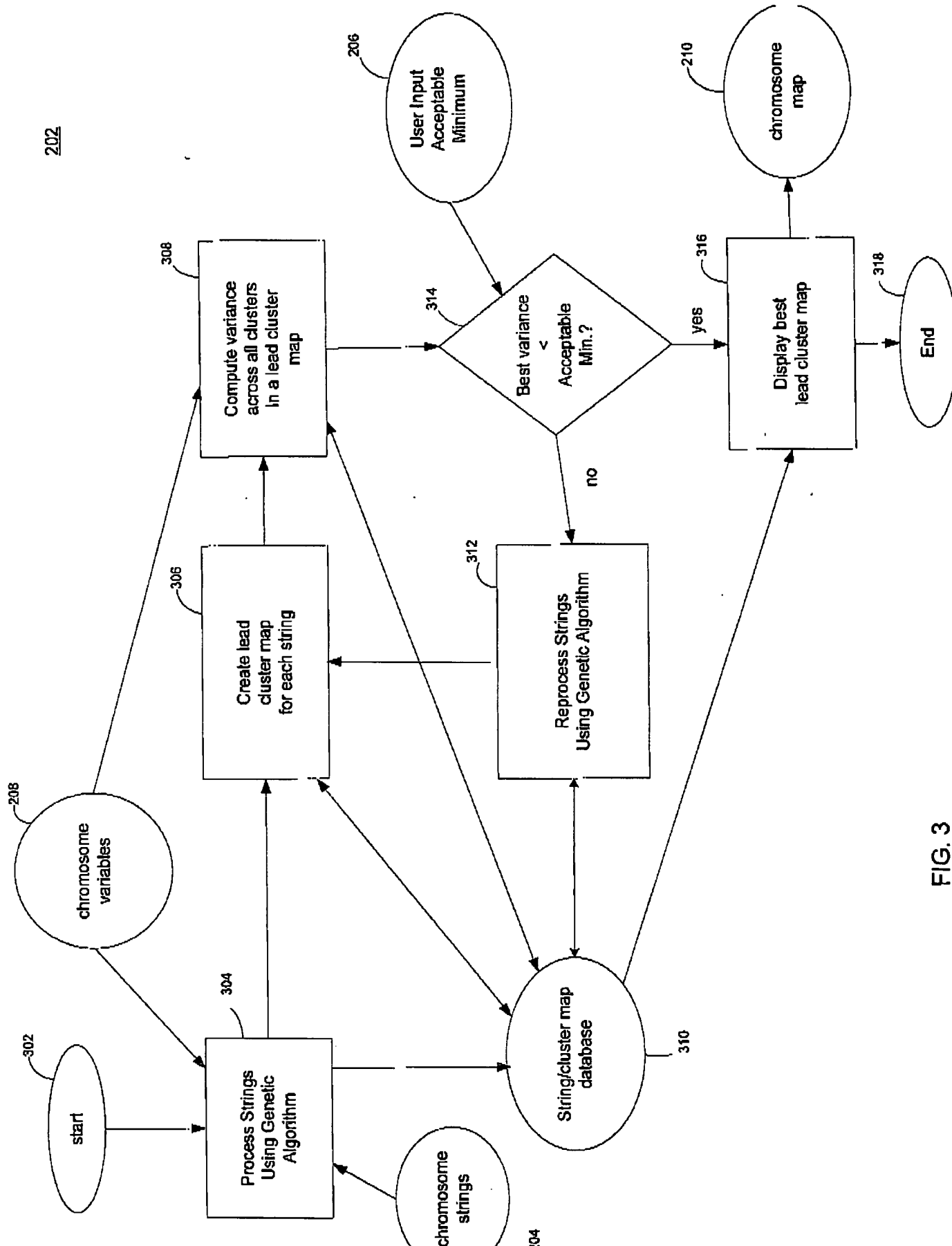


FIG. 3